DOCUMENT RESUME

ED 223 646 TM 820 727

AUTHOR Cohen, Peter A.
TITLE Synthesizing Research Results on Teacher Evaluation Using Meta-Analytic Procedures.
PUB DATE Aug 82
NOTE 19p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 23-27, 1982).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Course Evaluation; Curriculum Evaluation; Evaluation Methods; Feedback; Measures (Individuals); Program Effectiveness; Statistical Analysis; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; *Test Validity
IDENTIFIERS Meta Analysis

ABSTRACT

A meta-analysis of the literature findings on student evaluations of teaching are described. This research supports the validity of student ratings as measures of teaching effectiveness. Teachers whose students do well on achievement measures tend to receive higher instructional ratings than teachers whose students do poorly. Correlational effect sizes for the Skill and Structure dimensions were large; other rating dimensions, such as Rapport, Feedback, and Interaction, did not relate as well to student achievement. Correlational effect sizes for courses were remarkably stable under a variety of study conditions and methodological manipulations. These findings demonstrate that the relationship between ratings and achievement is stronger and more consistent than was previously thought. The advantages of meta-analytic techniques over traditional research reviews are discussed. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227040

Synthesizing Research Results on Teacher Evaluation Using Meta-analytic Procedures

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Peter A. Cohen

Dartmouth College

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. A. Cohen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Symposium paper presented at The Annual Meeting of The American Psychological Association, Washington, D.C., August 1982.

Introduction

This paper describes research in which I have used a technique called "meta-analysis" to synthesize findings in the literature on student evaluations of teaching. In general, student evaluations can serve three functions: 1) they provide feedback to teachers for the purpose of improving instruction; 2) they aid administrative evaluations of teacher effectiveness in decisions concerning pay increases, promotion, and tenure; and 3) they help students to select courses and instructors. The studies described in this paper address the first two of these functions. The first meta-analysis concerns the comparative effectiveness of student-rating feedback for improving college teaching. The second meta-analysis deals with the validity of student ratings, which is a crucial concern if ratings are to be used in promotion and tenure decisions.

Before summarizing the results from these two studies, I would like to discuss the rationale for using meta-analysis to synthesize research findings in the social sciences, and also to describe briefly what a meta-analysis actually is. One problem we have as social scientists is that our research literature tends to be extraordinarily inconsistent and complex. While researchers in the so-called "hard" sciences are able to reproduce the conditions of their experiments, for psychologists, findings from applied studies often vary in confusing ways from setting to setting. Although we look to research reviews to

summarize an area, it seems that reviewers seldom are able to find a pattern in the findings.

This confusion is particularly apparent in the area of teacher evaluation. There have been thousands of studies on course and instructor evaluation over the past decade, and researchers in the area have experienced considerable difficulty trying to stay current with this literature. But even with this flood of research studies, the conclusions drawn from the teacher evaluation literature are generated from relatively few studies. This state of affairs may be attributed, in part, to our current inadequacy with respect to analyzing results from a large number of diverse studies.

In his presidential address to the American Educational Research Association, Glass (1976) proposed a method for handling the difficulties posed by the diversity of findings in the social sciences. This approach, which he called "meta-analysis," involved the statistical analysis of a collection of results from individual studies for the purpose of integrating findings. By statistically summarizing outcomes from a large number of studies one can define relationships embedded in a complex literature more sharply. Reviewers who carry out meta-analyses first locate studies of an issue by clearly specified procedures. They then characterize the outcomes and features of these studies in quantitative or quasi-quantitative terms. Finally, meta-analysts use multivariate techniques to describe findings and relate characteristics of the studies to outcomes. By applying to a

collection of results the same objective methods that researchers use in analyzing results from an individual study, we can draw reliable and reproducible conclusions that have substantial advantages in precision when compared to traditional methods of research review.

Study 1: Effectiveness of Student-Rating Feedback for Improving Instruction

Traditionally, there has been substantial support for using student ratings as a source of feedback, but it has not been clear how effective ratings were in terms of improving instruction. A number of reviewers have noted conflicting findings--some studies revealing a significant influence of student-rating feedback while others showed no effects (e.g., Abrami, Leventhal, and Perry, 1979; McKeachie, 1979; Rotem and Glasman, 1979).

In initiating this meta-analysis on the effectiveness of student-rating feedback, I was interested basically in three questions. First, how effective is student-rating feedback in the typical comparative study? Second, is student-rating feedback especially effective for certain dimensions of instruction? And finally, under which conditions does student-rating feedback appear to be most effective? To conduct the meta-analysis I located 19 studies that met the following

criteria: A) the study had to be conducted in an actual college class; B) the study had to compare teaching improvement outcomes in two groups of instructors--those receiving mid-term student-rating feedback and those not receiving feedback; C) instructors being compared had to hold the major responsibility for teaching the class; and D) differences in initial teaching ability between groups of teachers had to be controlled. The 19 studies reported 24 separate comparisons involving mid-term student-rating feedback.[1]

Results

The first set of analyses examined the overall size and significance of effects of student-rating feedback. The feedback group received higher end-of-term overall ratings than the no feedback control group in 21 of the 24 comparisons (88%). Ten of the 24 comparisons reported statistically significant differences between feedback and no feedback groups, and, in each case, the difference favored the feedback group. As can be seen in Table 1, the mean effect size (ES) for the 24 comparisons was .41. Glass (1976) defined effect size as the difference between the means for two groups divided by the standard deviation of the control group. Therefore, the average effect of feedback was to raise end-of-term overall ratings by about 4/10's of a standard deviation. This size of effect can be considered moderate in size. Table 1 also lists results for seven dimensions of student ratings. For four of these dimensions--Skill, Rapport, Feedback

to Students, and Evaluation--mid-term student-rating feedback had a statistically significant impact on end-of-term ratings. Effect sizes for these dimensions, however, varied quite a bit.

Insert Table 1 about here

In terms of other outcomes, there were few comparisons upon which to base results. From the limited information it seems that students whose instructors received feedback rated their own learning higher than students whose instructors did not receive feedback. The comparisons on actual student achievement seem to verify students' perceptions. Similarly, students whose instructors received feedback were more positive in their attitudes toward the subject matter than students of control group instructors.

A second set of analyses determined whether comparisons reporting large effects differed systematically from those reporting small effects. The only significant effect here was for type of feedback. When feedback consisted of descriptive rating feedback (means, standard deviations, and frequency distributions of rating items) there was a small teaching improvement effect for instructors receiving feedback (ES=.19). If rating feedback was augmented in some fashion the effect was a little larger (ES=.44). But when rating feedback was accompanied by some sort of consultation, the teaching improvement effect was quite substantial (ES=.84).

Summary

The trends from this body of research are relatively consistent, and therefore we can draw some overall conclusions concerning the effectiveness of student-rating feedback. Clearly, student ratings can be a useful feedback device, especially when the instructor, in conjunction with a consultant, uses them to devise teaching strategies. It is interesting to note, however, that rating feedback, even when mediated by a consultant, does not lead to substantial improvement for all dimensions of teaching. Items from the Skill, Interaction, and Evaluation rating dimensions show the most change from mid- to end-of-term. These types of items address more specific teaching behaviors, which are more easily modified than general course and instructor characteristics.

Study 2: The Relationship between Student Ratings of Instruction and Student Achievement

This second study concerns the validity of student ratings. A standard procedure for conducting student rating validity studies has been to correlate mean instructor ratings with mean class achievement in a multisection course, using a common rating instrument and achievement measure. Although past reviewers (Kulik & Kulik, 1974; McKeachie, 1979) have acknowledged that in general there seem to be small to moderate correlations between ratings and learning, some people have pointed to studies showing strong positive correlations (e.g., Centra, 1977; Frey, 1973), while others have cited studies showing equally strong negative

correlations (e.g., Rodin & Rodin, 1972), depending on the view they are espousing on student ratings.

To be included in this meta-analysis a study had to meet three basic criteria: A) it had to provide data from actual college classes; B) the unit of analysis had to be the class rather than the individual student; and C) data had to be based on a multisection course with a common achievement measure and rating instrument used for all sections. Using these criteria I located 41 studies reporting on 68 multisection courses.[2] I was interested in determining several outcomes of the studies. Primarily, I wanted to know the relationship between student achievement and overall ratings--overall course ratings and overall instructor ratings--because these ratings are most often used in evaluating teaching effectiveness. I also wanted to determine the relationships between student achievement and more specific rating dimensions commonly found in factor-analytic studies of student ratings.

Results

Table 2 shows the mean rating/achievement correlations, the number of multisection courses on which the correlations are based, and the 95% confidence intervals on the mean population correlations. For 20 of the 22 courses (91%) reporting overall course rating/achievement correlations, the correlation was positive, and in 11 of these courses the correlation was statistically significant. The average rating/achievement correlation here was .47, quite a large correlational effect. For 59 of 67 courses (88%), there was a positive correlation

between overall instructor rating and student achievement. The correlation coefficient reached statistical significance for 30 of these courses (there was one significant negative correlation). The average correlation in the 67 courses was .43.

Insert Table 2 about here

Table 2 also shows rating/achievement correlations for other dimensions of teaching. While large effect sizes are found for the Skill and Structure dimensions, other dimensions such as Rapport, Interaction, Feedback, and Evaluation show more modest effects. The Course Difficulty dimension shows no relationship with student achievement. Finally, students' self-ratings of their learning correlate quite highly with student achievement.

Another goal in this meta-analysis was to explain the variation in effects among the different studies. To characterize studies more precisely, 20 variables were defined describing methodological features of the studies, conditions under which the studies took place, and publication features of the studies. Three of these variables were significantly related with effect size: control for bias in evaluating achievement; time at which the ratings were administered; and instructor experience. A hierarchical regression analysis showed that when achievement tests were graded by students' own instructors, the correlation between overall instructor ratings and student

achievement averaged .15; the correlation was .52 when an external grader was used. In terms of the time at which the ratings were administered, the average correlation was much higher when students knew their final grades (.85) than when they did not know their final grades (.38). For graduate student instructors, the correlations between overall instructor rating and student achievement averaged .34, while for full-time faculty the correlation was .48.

Summary

This meta-analysis provides support for the validity of student ratings as measures of teaching effectiveness. Teachers whose students do well on achievement measures tend to receive higher instructional ratings than teachers whose students do poorly. It is evident that certain aspects of teaching, as measured by student ratings, are more related to learning than are others. Correlational effect sizes for the Skill and Structure dimensions were large; other rating dimensions, such as Rapport, Feedback, and Interaction, did not relate as well to student achievement. The results also show that characteristics of the study do not, in general, have much of an impact on study outcomes. Correlational effect sizes for courses were remarkably stable under a variety of study conditions and methodological manipulations. While these findings reinforce conclusions reached by earlier reviewers, they demonstrate that the relationship between ratings and achievement is stronger and more consistent than was previously thought.

Conclusions

The use of meta-analytic techniques for synthesizing findings has several advantages over traditional research reviews. First of all, most meta-analysts follow clearly specified procedures for locating and selecting relevant studies. When the number of studies in an area run up into the hundreds, reviewers can often find what they want to in a collection of results. The use of objective procedures makes a meta-analysis more like an empirical research study and puts some constraints on a reviewer's fancy. Secondly, meta-analysis makes it possible to reach more exact conclusions about a research literature. In addition to a narrative description of a literature, reviewers using meta-analysis are able to calculate the size and significance of overall effects in an area. Moreover, meta-analysis helps determine the generalizability of findings from one setting to another, and, specifically, whether certain study characteristic variables systematically relate to size of effect. Finally, meta-analysis shows investigators what areas in a research literature are in need of future study.

References

Abrami, P. C., Leventhal, L., & Perry, R. P. Can feedback from student ratings help to improve college teaching? Proceedings of the 5th International Conference on Improving University Teaching. London, 1979.

Centra, J. A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

Cohen, P. A. Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. Research in Higher Education, 1980, 13, 321-341.

Cohen, P. A. Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 1981, 51, 281-309. (a)

Cohen, P. A. Using student ratings to improve instruction: A synthesis of research findings. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981. (ERI Document Reproduction Service No. ED 200 647) (b)

Frey, P. W. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

Glass, G. V Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Kulik, J. A., & Kulik, C-L. C. Student ratings of instruction. Teaching of Psychology, 1974, 1, 51-57.

McKeachie, W. J. Student ratings of faculty: A reprise. Academe, 1979, 65, 384-397.

Rodin, M., & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1,164-1,166.

Rotem, A., & Glasman, N. S. On the effectiveness of students' evaluative feedback to university instructors. Review of Educational Research, 1979, 49, 497-511.

Footnotes

[1]For a complete description of the search procedures and a list of studies used in this meta-analysis see Cohen (1980) and Cohen (1981b).

[2]For a complete description of the search procedures and a list of studies used in this meta-analysis see Cohen (1981a).

Table 1

MEAN RATINGS AND EFFECT SIZES
FOR MAJOR OUTCOME MEASURES

Outcome	Number of Comparisons	Means		t	Mean Effect Size
		Feedback	Control		
Overall Rating	24	3.89	3.75	4.03***	.41
Skill	17	3.96	3.80	5.19***	.51
Rapport	13	4.06	4.00	2.22*	.13
Structure	12	3.76	3.61	1.68	.25
Difficulty	6	2.44	2.40	.95	.09
Interaction	13	4.06	3.91	2.04	.37
Feedback to Students	7	3.65	3.49	2.93*	.33
Evaluation	6	4.02	3.84	3.00*	.40
Student Self-rating of Learning	5	3.88	3.76	5.66**	.36
Attitude Toward Subject	4	3.30	3.15	4.95*	.49
Student Achievement	4				.39

*p < .05
**p < .01
***p < .001

Table 2

MEAN RATING/ACHIEVEMENT
CORRELATIONAL EFFECT SIZES

Rating Dimension	N	Mean Correlation	95% Confidence Interval
Overall Course	22	0.47	0.09, 0.73
Overall Instructor	67	0.43	0.21, 0.61
Skill	40	0.50	0.23, 0.70
Rapport	28	0.31	-0.07, 0.61
Structure	27	0.47	0.11, 0.72
Difficulty	24	-0.02	-0.42, 0.39
Interaction	14	0.22	-0.36, 0.67
Feedback	5	0.31	-0.79, 0.94
Evaluation	25	0.23	-0.18, 0.58
Student Progress	14	0.47	-0.08, 0.80